Available online at www.sciencedirect.com
ScienceDirect
Procedia Economics and Finance 33 (2015) 277 – 286
Procedia Economics and Finance
www.elsevier.com/locate/procedia

7th International Conference, The Economies of Balkan and Eastern Europe Countries in the changed world, EBEEC 2015, May 8-10, 2015

Using Big Data in the Academic Environment

Banica Logica[a*], Radulescu Magdalena[a]

[a]*Faculty of Economics, University of Pitesti, Targu din Vale Street, no. 1, 110040 Pitesti, Romania*

Abstract

Massive amounts of data are collected across social media sites, mobile communications, business environments and institutions. In order to efficiently analyze this large quantity of raw data, the concept of Big Data was introduced. This new concept is expected to help education in the near future, by changing the way we approach the e-learning process, by encouraging the interaction between students and teachers, by allowing the fulfilment of the individual requirements and goals of learners.
The paper discusses aspects regarding the evolution of Big Data technologies, the way of applying them to e-Learning and their influence on the academic environment. Also, we have designed a three-step system architecture for a consortium of universities, based on actual software solutions, having the purpose to analyze, organize and access huge data sets in the Cloud environment. We focused our research on exploring unstructured data using the graphical Gephi tool.

Peer-review under responsibility of Department of Accountancy and Finance, Eastern Macedonia and Thrace Institute of Technology

Keywords: Big data; E-learning; Academic environment; Cloud Computing.
JEL classification codes: C82, I21, D83.

1. Introduction

In last years the IT world is facing with a massive increase in the produced data volume, mainly due to the Internet services, leading to the redefining the term database. The new concept used for the description and

* Corresponding author. Tel. +40-745-227-774
E-mail address: olga.banica@upit.ro

Peer-review under responsibility of Department of Accountancy and Finance, Eastern Macedonia and Thrace Institute of Technology
doi:10.1016/S2212-5671(15)01712-8

organization of enormous quantities of data, structured or unstructured, provided by companies, institutions and by social media environments is Big Data.

The first definition of term originating in 1997, and was introduced by two NASA researchers, Michael Cox and David Ellsworth (1997); in 1998, John R. Mashey, a researcher from Silicon Graphics Inc. (SGI) used this concept (1998), and after one year, Bryson et al. published a paper concerning Big Data in the Communications of the Association for Computing Machinery (ACM) (1999).

Starting from 2009, research gave birth to hardware and software solutions, and merged with other technologies that support the entire ecosystem in order to achieve the desired goals. This way, the Cloud Computing environment provided the resources required to store and access important data volumes, and thus facilitated and developed the symbiosis of them. These emerging technologies became a foundation for the e-Learning industry and offered an opportunity to help higher education in the near future, by changing the way to approach e-learning process, by encouraging the interaction between students and teachers, by allowing them to follow the individual needs and performances of learners.

The paper aims at analysing these aspects, especially the applications of Big Data, supported by Cloud Computing, in the e-Learning process. We designed a three-step model architecture for a university e-learning system, based on actual software solutions, having the purpose to achieve, organize and access huge data sets in cloud environment.

Our work includes three sections and a Conclusions part. Section 2 presents the three concepts involved (Big Data, Cloud Computing and e-Learning System), summarizing the state-of-the-art, and investigates several methods to store, filter and process a large volumes of data, with the help of commercially-available software solutions. Also, we briefly presented the characteristics of Gephi software. Section 3 specifies the methodology used, by presenting software platforms necessary for each of the three levels of the designed architectural model. Section 4 includes our solution for a system that is able to accommodate Big Learning Data in a Public Cloud environment. The main concluding remarks close the paper, and suggest ways of improving our future research activity in this domain.

2. Literature Review

2.1. Big Data concept

The appropriate definition of the concept, as seen by the authors of this paper is the following: “Big data is a massive collection of shareable data originating from any kind of private or public digital sources, which represents on its own a source for ongoing discovery, analysis, and Business Intelligence and Forecasting”, according to Banica et al (2014).

The most important volume of data is provided by social media sites and mobile networks, but the percent of useful information is reduced, in comparison with other categories of data sources which are more valuable, such as financial and governmental institutions, education institutions and the business environment.

Big data, in the context of e-Learning systems (also called Big Learning Data), consists in the information sources (courses, modules, experiments etc.) created by the teachers, but especially in data coming from the learners (students) throughout the education process, collected by the Learning Management Systems, social networks, multimedia, as they were defined by the organization or the professionals.

Oracle described Big Data by four keys characteristics (the four Vs): Volume, Velocity, Variety and Value as point out Dijcks (2013) and Briggs (2014). By adding these features to the Big Learning Data, we will further describe the content and the importance of every key characteristic, in the four Vs approach:

- Volume: the size of the data. It is difficult to define the limits for Big Data, so this is a very relative aspect for every domain of application, also for the education field. In our opinion, even if data originates from thousands of students in one university, we consider that Big Data term may be used if several higher education institutions collaborate in the information exchange and might bring together learning data and learners.
- Velocity: the increasing flows of data need hardware and communication equipment able to carry more and more information, and software solutions to process them as fast as possible; Big Learning Data must ensure for

students and teachers the quick access to the information, needed in educational process; for example to correct a wrong answer into an assessment exam or to allow teachers to make adjustments to the content of course during the class, or to answer to the students questions in real time.

- Variety: Big Data is a combination of all types of formats, unstructured and multi-structured. Therefore Big Learning Data collects, analyzes and provides the information with different backgrounds to ensure better learning resources; the focus is on handling them, so there should be no various behaviors or performances.
- Value: concerns the scientific value or the commercial value of Big Data. So, if for enterprises it is important to use data originating from social media in combination with internal data in order to develop their business, for educational environment is more important the degree of innovation. The target of Big Learning Data is to obtain a high level of education and knowledge, and to develop projects in research domains, that lead, as a consequence, to new solutions in all areas (economic, financial, health, education and social).

2.2. Meting Big Data with e-Learning in Cloud environment

The problem of hardware and software resources needed to store huge volumes of data may be solved by using Cloud Computing technology. Not only the business environment is interested in collecting information from unconventional data sources, but also government agencies, higher education institutions and other organizations analyze and extract meaningful insights from this labyrinth of data, underlines Ferkoun (2014), be it security related, behavioral patterns of consumers or feedback to the e-Learning courses.

It is difficult to suppose that ordinary businesses and institutions could afford such advanced technological resources; therefore we believe that the development of Big Data relies mostly on Public and Hybrid Cloud implementations. There is a powerful symbiosis between these two technologies, considering that any Cloud Computing implementation includes a high-capacity storage solution and any Big Data platform needs to collect, analyze and process large volumes of data, from multiple sources and in various types.

In a hierarchical structure, the base could be the Cloud, which would offer the resources, Big Data would come in the middle, which would be responsible for data organization and processing, and at the top we may develop new e-Learning industry opportunities.

We presented in several papers Cloud-based Big Data scenarios and we emphasized that Hybrid Clouds are often the preferred option for the institutions and companies, which may use Private Clouds to manage internal structured data, while Public Clouds allow the storage of volumes of external data or archives (Big Data).

Many powerful corporations like Google, IBM, Sun, Amazon, Cisco, Intel, and Oracle have invested in a wide range of cloud-based solutions, which confirms that this is a technology that they rely on, and from whom there are great expectations. In such a competitive environment are taken into account many aspects: the storage capacity, the security of hosted data, the services provided, but also the subscription cost.

Cloud Computing has also several drawbacks, some of them of major importance, according to Banica et al. (2014):

- data security risks – secured access policies are required in order to keep unauthorized users away from the business data;
- data loss challenge – all databases are required to implement automatic backup and transaction-based queries, thus mitigating the chance to affect service quality;
- system unavailability – network outages and OS crashes can negatively affect the performance of the solution, and redundant architectures must be implemented by providers.

Banica and Burtescu (2014) argue that some of the security problems were solved, not entirely, but at a reasonable level by new ways to ensure protection from unauthorized access to the Cloud by layered security approach (different sets of user privileges, grouped in access roles), by firewall policies, with powerful rule sets for Internet and Intranet access, by usage of cryptography for the data, or by using smart solutions for traffic filtering with automated alerting.

Big Data in the Cloud is a powerful platform for the e-Learning world, especially for the higher education, in ways that we will try to emphasize in the following section.

2.3. Some motivations of introducing Big Data in e-Learning

We consider that the preferred option for the universities is, as for the business environment, the Hybrid Cloud solution, which may use Private Clouds for the learning management systems (LMS), while Public Clouds is dedicated to storing and processing Big Data, consisting mainly in unstructured data from students, via social networks and other media.

Universities all over the world are using learning management systems (LMS), based on integrated collaborative software platforms. Applications like wikis, chat rooms and blogs enable teachers to continuously observe and check the progress of students, and students to communicate more efficiently among them and with their teachers, in order to faster and better evolve in a knowledge field.

Resource sharing and exchanges of ideas are the perfect support for a teacher that wants to know the level of knowledge of the students, about the topics proposed for study. A discussion of the educational potential of collaborative software needs to be started from the point of view of the involved groups of students, on one side, and from the one of the learning professionals, on the other side, as pointed out by Banica (2014).

We have asked ourselves, as probably many other university professors did, if this model is working for many universities, as teachers give to their students some learning items, called Learning Objects, and may cooperate with them maintaining a professional relationship, by building weblogs and wikis for courses or projects and the students could have an appropriate way to communicate with the teachers, why is necessary the migration of an LMS to a Big Learning Data in Cloud environment? Or, more directly, how Big Data could bring performance to e-Learning process?

The answer is not so simple, taking account of the evolving volume of information, the liberty of expression and the candor to be found in social media and the differences between university budgets – compared to the strong need for equal opportunities for the students around the world.

This project is not aimed at the students of a single university, but to any given consortium of many educational institutions, that could enrich the knowledge of the learners and open the way for comparative analyzes, and thus overcome the lack of financing the Cloud-powered Big Data from local resources.

There are many advantages of Big Data for the e-Learning process:

- from teacher's angle - the opportunity to understand the real patterns of their students, to assess the current level of their knowledge, as Briggs (2014) says, to determine which parts were too easy or too difficult and improve the content by enabling them to personalize courses;
- from the point of view of students – ensuring a rich communication and offering endless learning opportunities.

Big Learning Data based on Cloud computing is a new concept and its application for higher education is just starting, but we could say that its deployment involves the following important stages, according to Banica et al. (2013):

- ensuring a powerful infrastructure, including the hardware and software computing resources required for e-learning activities;
- finding a unifying solution for the Learning Objects representation, given that most universities have their own Learning Management System, implemented in private cloud, but it is a different one for each institution;
- keeping the privately-owned cloud to ensure the confidentiality of scholar records, teaching personnel data and research projects;
- the access to the educational content and entirely use the collaboration tools for the students and teachers from the universities involved in the e-Learning Cloud project ;
- building open-source cluster architectures for gathering and processing the unstructured information.

Not always the amount of data may be the cause of introducing Big Data in education, the main problem could be their heterogeneity, most of available data being unstructured, that prevents the usage of the classical relational databases. Therefore, the IT world launched a new wave of technologies, capable to solve the problem, such as Hadoop and Spark software.

High-profile vendors such as Amazon Web Services, Google, Microsoft, Rackspace and IBM offer cloud-based Hadoop and NoSQL database platforms for Big Data, underlines Vaughan (2014). Due to the services that run on these platforms, and taking advantage of the reduced costs and increased flexibility, Big Data on Cloud computing is the first choice for business companies.

That cannot be said about public institutions or educational organizations, which have a slower adoption rate and mention the lack of security as the main cause. But the motivation to use the Cloud is far more powerful for the universities, which can benefit from borderless access to knowledge and information exchange between learners and educators, and empower scientific research.

3. Methodology

In this section we will briefly present the new type of database - NoSQL, used for storing Big Data, the software that allows for massively parallel computing – Hadoop, and the Social network analysis (SNA) software Gephi.

NoSQL databases (or "Not Only SQL") encompass a wide variety of unstructured data described by several models: key value stores, graph, and document data. NoSQL data may be implemented in a distributed architecture, based on multiple nodes, able to process and store Big Data.

In the dedicated literature, there are four types of NoSQL databases, each of them having specific attributes, such as they are mentioned by Mohamed et al. (2014):

- Key-value store (KVS) – uses a hash table of keys and values for designing databases; in this table a unique key exists, and a pointer to each record of data; the key-value model is inefficient for querying and updating part of a value. Example of key-value databases: Oracle BDB, Amazon SimpleDB
- Document – represents the next level of Key-value type, where data is a collection of key value pairs, compressed as a document in different format standards, such as XML or JSON. It is a complex category of storage that enables data querying more efficiently. Examples of document databases: CouchDB, MongoDB;
- Column – refers to a database structure similar to the standard relational databases, data being stored as sets of columns and rows. The columns are logically grouped into column families. Column category databases are recommended to be used when the number of write operations exceeds reads, for example in logging. Examples of Column databases: Cassandra, HBase;
- Graph – designs the structures where data may be represented as a graph with interlinked elements. Instead the rigid structure of SQL, based on tables of rows and columns, a flexible graph model with edges and nodes is used, scanning across multiple machines. In this category, social networking and maps are the main applications. Examples of Graph databases: Neo4J, InfoGrid, Infinite Graph.

Choosing a data model for NoSQL solution depends on technical differences and working features, especially on keeping data consistency and a very fast retrieval.

The evaluation of NoSQL implementations takes into account: the storage capacity, the ability to use memory efficiently, the support for deployment on virtual machines, and the Cloud, but also the execution time for different operations, as indicated by Henschen (2014).

According to Baby (2014), a researcher from IGI Global's InfoSci-Dictionary, *Hadoop* is „an open source, Java-based programming framework that supports the processing of large data sets for scalable and distributed computing environment".

Essential to the effectiveness of this software is to do the processing in proximity to the location where data is stored and not to bring the data to the computation units, preventing unnecessary network transfers, point out the scientists from *Yahoo* Developer Network (2007). Its parallel-processing capability is better used when deployed in the Cloud, because large amounts of data stored in the cloud can be processed, queried and analyzed at high speeds.

Hadoop can be installed on any of operating system (OS) families (Linux, Unix, Windows, Mac OS) and can be run on a single node or on multi-node cluster. A Hadoop distribution includes two core parts: the storage component, Hadoop Distributed File System (HDFS) and the processing component, the MapReduce engine. The base Hadoop framework also contains a module for libraries and utilities (Hadoop Common) and a module responsible for managing and scheduling cluster resources (Hadoop YARN). HDFS splits large data files into blocks which are distributed amongst the nodes of the cluster and are managed by them and also are replicated across several machines, for security reasons, according to Frank Lo (2014). MapReduce is the key component that Hadoop uses to transfer blocks around a cluster, so that operations can be run in parallel on different nodes and data is to be processed locally.

Hadoop has the advantage that it can be used in the Cloud environment and supports distributed data processing for Big Data across clusters. Henschen (2014) suggests that today, Hadoop is the preferred solution for Big Data architectures and practically, the two technologies have become synonymous.

The most widespread solution is the open source Apache Hadoop distribution, but there are powerful software corporations, such as IBM and Oracle, that have their own distributions. Also, Yahoo and Facebook seem to have the largest Hadoop clusters in the world. This technology deployed in Cloud environments is offered as a service by companies, such as Microsoft, Amazon, and Google.

Figure 1 shows the functional structure of Hadoop, the stages that data passes through from acquisition to storage in the NoSQL database.

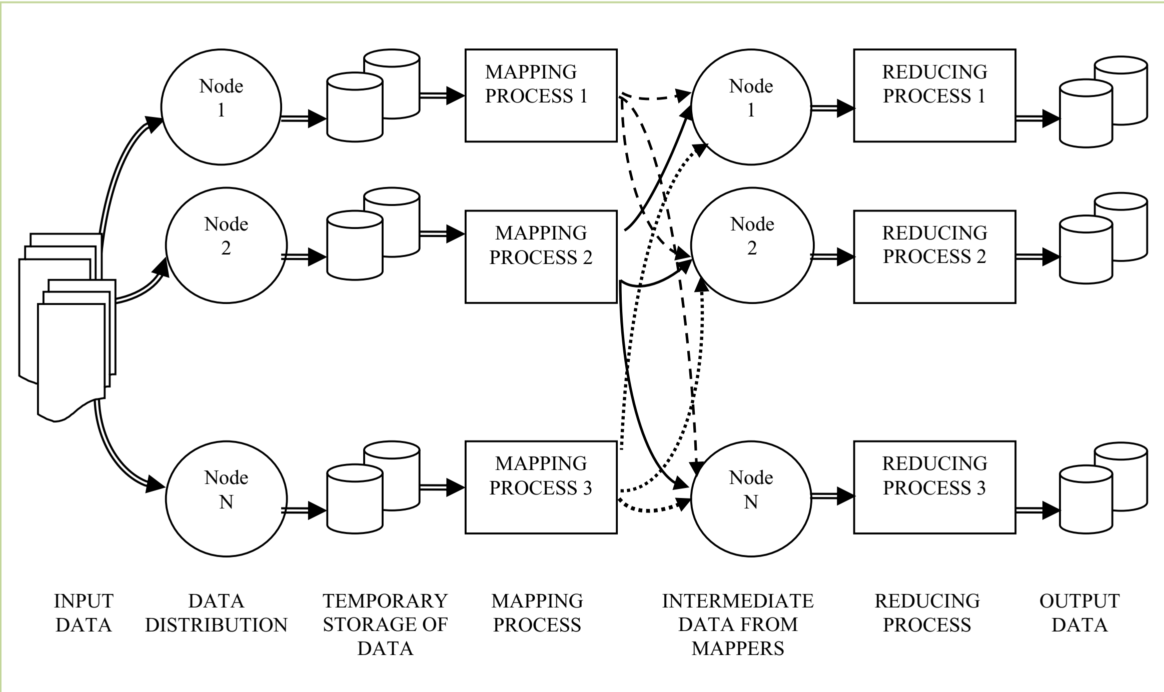

Fig.1. Big Data Processing with Hadoop (Source: Banica et al., 2014)

An important question that our study should answer is following: how do existing Learning Management Systems interact with Hadoop?

Obviously, the new architecture for the Learning Management Systems is designed to complement and to extend the existing systems, and not replace them. As the previous versions of LMSs are based-on RDBMSs and Hadoop has native support for extracting data over JDBC, one solution is dumping the entire database to HDFS and making updates, according to Bisciglia (2009). Another option could be the transfer of LMSs structured data into a consolidated Data Warehouse. There are several tools to do these operations, such as Apache Sqoop or ETL (Extract, Transform, Load) process, which can collect data originated from external databases.

After the output data storage could be applied conversion and filtering operations and then conducted advanced analysis, having different goals: finding correlations across multiple data sources, predicting an entity behavior, or analyzing social networks.

In our study, we suggest that the system will apply a new filter in order to search for certain keywords, and store the results in .csv, .gml, .gdf, .gefx files, or even spreadsheets, from which they can be further explored using social network analysis (SNA) software, such as Gephi. In a Report of the International Institute for Sustainable Development, published on 2012, Ryan and Creech (2012) mentioned that „Social network analysis software is used to identify, represent, analyze, visualize, or simulate nodes (e.g. agents, organizations, or knowledge) and edges (relationships) from various types of input data (relational and non-relational), including mathematical models of social networks."

In addition to providing a networked visualization, such software generates metrics, identifies subgroups in a network, clusters of actors or individuals, or emphasizes isolated nodes of the network.

Krebs (2013) considers centrality a commonly used measure, which refers to the importance of a node into the network and the hierarchy of the entire network. Another important measure in SNA is network density. This measure is useful for assessing the overall relationships within a network of *n* nodes.

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, up to 50,000 nodes and 1,000,000 edges. In opinion of Bastian et al. (2009) this is an application that implements the most frequently used algorithms in descriptive statistics for networks.

After the graph was built, controls can be applied in order to select nodes and/or edges, to view their implications on the network structure or to measure average accesses, the groups with most frequent accesses. The graph can be undirected, representing only symmetric relations, directed, for asymmetric and symmetric relations and weight, representing intensities, distances or costs of relations. Gephi works with imported files from .csv, .gml, .gdf and .gefx format, which can be achieved with software converters (e.g. Facebook or Twitter to .gdf files) from unstructured data, by applying an algorithm that transforms key-value words on nodes and their connections on edges. The Gephi tool was successfully used to analyze a personal page of the Facebook network.

In our case, we will analyze social network data from the point of view of a teacher, in relation with their students, and the interconnections among the students themselves.

4. A Model for Big Learning Data on Cloud Architecture

In this section we will introduce a three-layered Cloud-enabled Big Data architecture for e-Learning, based on a Hadoop cluster which belongs to a given consortium of universities.

The main task is not Hadoop itself, as it is offered as SaaS by many cloud providers, but integrating it with the existing LMSs that the universities usually already own. Thus, our model refers to a unified architecture, using Hadoop as a data integration platform.

The power of Big Data consists in aggregating flows from social media, such as course information and availability, services, project collaboration and all gathered feedback, for the education environment.

For the moment, teachers can continue their activity without taking Big Data into account, but if they want relevant insights on their real efficiency and the progress of their students, they need to integrate this kind of solutions. Thus, Big Learning Data is not an unstoppable information flow that affect operational applications, but a chance to refine the educational process, to adapt to the new requirements. The key is not sheer information volume, but it is complexity and diversity.

So, the LMS from each university of the consortium should accept the transfer of their Learning Objects, Student Information System and Teacher Information System into the Data Warehouse. Data Warehousing (DW) is a method of organizing structured data, built upon a relational database, therefore it needs to have all the information consistent, organized and standardized. But, a DW cannot capture an important data segment (clickstream logs, sensor and location data from mobile devices, customer emails and chat transcripts) and this is the point where Big Data systems prove their importance, allowing to analyse and extract educational value from this unstructured information.

There are two scenarios of the integration between Hadoop and DW, according to Dumitru (2011):

- Using ETL process (*Extract, Transform, Load)* to obtain Data Warehouse from heterogeneous data collected by Hadoop and then applying advanced analytics to Data Warehouse;

- Considering Hadoop as a data integration platform, it is able to collect data from all type of data sources, then to process it in order to make the data suitable for analysis.

The model presented in intended to work on both scenarios, taking into account that its target is the educational world and sometime is needed a subset of the initial data (first version) and, most of the time, is needed access to the actual data, not only to a subset of it (second version). Also, we focused on the third level, which begins after the data flow was processed based on Map and Reduce functions.

For both SQL and NoSQL databases, the proposed solution involves a classification software that could differentiate domains and subdomains, a categorization that is standard for Data Warehouses, and that must be generated for unconventional data representations. For example, a criterion could be the counting of word occurrences and comparing this number with Domain Dictionaries. A metadata level is added, in order to direct data to specific storage locations (SQL database or unstructured datasets). The identification of trends, patterns or clusterization models is based on keywords, which trigger a parallel processing of the data stored for the required domain/subdomain, using both NoSQL and SQL systems, as we already claimed in another paper, Banica et al. (2014).

In figure 2 is shown the Big Learning Data model proposed for a consortium of Romanian universities, in a three-level structure:

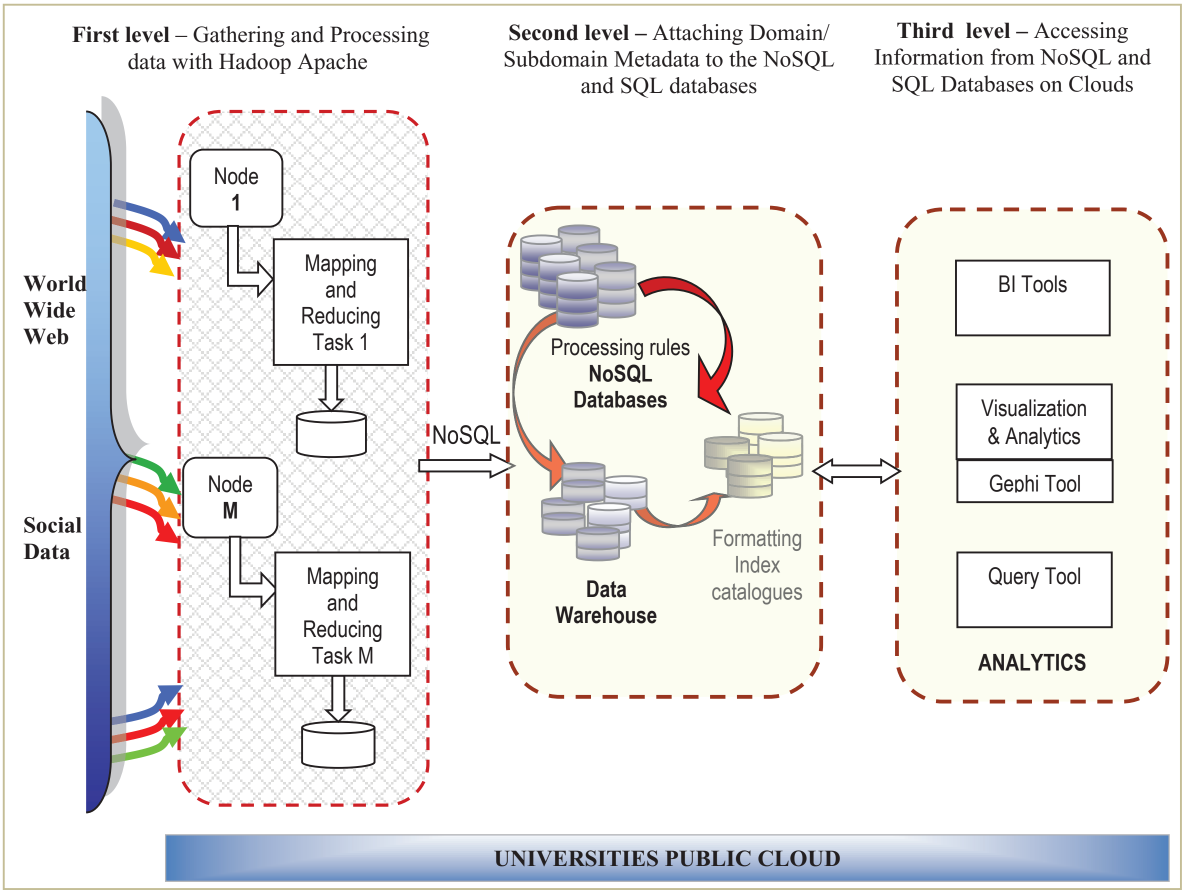

Fig.2. A model for Big Learning Data on Cloud architecture

- Gathering and Processing of all kinds of data (structured and unstructured) with Apache Hadoop, interesting data for the universities involved in the project;
- Attaching them Domain/Subdomain Metadata (different treatment for NoSQL and SQL databases);
- Processing filtered subsets of NoSQL Databases with SNA software, such as the Gephi tool, and finding a pattern, a trend as the response to the search requirements.

The *first level* is designated for collecting any type of data based on software tools such as Apache Flume (unstructured data), Apache Sqoop (structured data) and ETL (structured data) and processing them using Hadoop cluster. The most important volumes of data are the one originating from social media, containing relatively low amounts of useful information, compared to a smaller volume of data from educational institutions, containing a bigger percent of useful information. This level can be based on Public Clouds or Grid computing environments, but we find the Public Cloud approach more suitable for the academic budgets.

At the *second level* we propose a new software layer, having the goal to realize a classification of the data stored in the Hadoop cluster nodes into domains and subdomains, using dictionaries and the rule set that inserts specific metadata. For structured data, there are a several tools for metadata management into the Data Warehouse, such as MetaStage. The core components of Hadoop itself have no special capabilities for cataloging, indexing, or querying structured data.

The *third level* is focused on further ways to access the information. We consider that an efficient method to query this kind of data would be to build Index Catalogues, by using accompanying metadata and place the original information into separate storage spaces according to its type. Our model involves also the existence of a search engine for the Index Catalogues, based on keywords and data types.

Also, we are interested in how unstructured data can be processed in order to discover useful insights for teachers and students. In NoSQL databases, a search is performed using Index Catalogues, by mentioning the search terms (such as: university, course title and teacher name) and results are stored in Gephi-compatible formats (such as .csv, spreadsheets, .gml, .gdf etc).

The Gephi node-table is built, starting from the main node (teacher name) and followed by lower-level nodes (students enrolled in the course or interested on that topic). The next step consists in creating the connection table, where all information related directly or indirectly to the teacher are placed.

5. Conclusions and Future Work

The traditional e-Learning architecture is obsolete and a new data model, that integrates Hadoop to the existing systems, is emerging in the IT world.

The entire solution described is efficient, because activities are separated on levels and resources, the traffic is managed by Hadoop in Clouds, and the analysis is able to add graphic representations to other types of results.

Considering that Big Data in the Cloud environment solutions, promoted by the biggest software companies, are performant, but also expensive, we will recommend a unified learning management system based on open-source software. Thus, the evolution of open-source products for this domain will allow universities to benefit from this new trend that empowers today's education.

In our future research we intent to implement a multiple node Hadoop cluster and evaluate its performance working with structured data from our university LMS and unstructured data from Social media.

We will expand this project by approaching the second level, by assessing the available tools that allow adding metadata to the relational and No-SQL databases, and creating index catalogues for interest domains in order to improve retrieval.

Also we intend to analyse the performance of several analytic tools (Level 3 of the proposed model), testing them against workloads with query and graphic interpretation of Big Learning Data.

References

Baby, N., Pethuru, R., IGI Global, 2014, Big Data Computing and the Reference Architecture, What is Hadoop, http://www.igi-global.com/dictionary/hadoop/12699

Banica, L., Burtescu, E., Stefan, C., 2014, Advanced Security Models for Cloud Infrastructures, Journal of Emerging Trends in Computing and Information Sciences, Vol. 5, No. 6, pp. 484-491

Banica, L., 2014, Different Hype Cycle Viewpoints for an e-learning System, Journal of Research & Method in Education, Vol. 4, Issue 5, pp.88-95, 2014

Banica, L., Stefan, C., Rosca, D. & Enescu, F., 2013, Moving from Learning Management Systems to the e-Learning Cloud, *AWERProcedia Information Technology & Computer Science*. pp 865-874, www.awer-center.org/pitcs.Bisciglia, C., 2009, 5 Common Questions About Apache Hadoop, available at http://blog.cloudera.com/blog/2009/05/5-common-questions-about-hadoop/

Banica, L., Paun, V., Stefan, C., 2014, Big Data leverages Cloud Computing opportunities, International Journal of Computers & Technology, Volume 13, No.12, http://cirworld.org/journals/index.php/ijct/article/view/3036

Bastian M., Heymann S., Jacomy M., 2009, Gephi: an open source software for exploring and manipulating networks, *International AAAI Conference on Weblogs and Social Media*, San Jose, USA

Briggs, S., 2014, Big Data in Education: Big Potential or Big Mistake? http://www.opencolleges.edu.au/informed/features/big-data-big-potential-or-big-mistake/

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D., and Haimes, R., 1999, Visually exploring gigabyte data sets in real time, http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Cox, M., Ellsworth, D., 1997, Application-Controlled Demand Paging for Out-of-Core Visualization, *Proceedings of the 8th IEEE Visualization '97 Conference*, available at http://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/ parallel_volviz/paging_outofcore_viz97.pdf

Dijcks, J., 2013, Big Data for Enterprise, http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Dumitru, A., 2011, Hadoop - Enterprise Data Warehouse Data Flow Analysis and Optimization, OSCON Open Source Convention, Portland, available at http://www.oscon.com/oscon2011/public/schedule/detail/21348

Ferkoun, M., 2014, Cloud computing and big data: An ideal combination, available at http://thoughtsoncloud.com/2014/02/cloud-computing-and-big-data-an-ideal-combination/

Frank Lo, 2014, Big Data Technology, available at https://datajobs.com/what-is-hadoop-and-nosql

Henschen, D., 2014, 16 Top Big Data Analytics Platforms, available at http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

Krebs, V., 2013, Social Network Analysis, A Brief Introduction, available at http://www.orgnet.com/sna.html

Mashey, R., J., 1998, Big Data and the Next Big Wave of InfraStress, Usenix conference, available at http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

Mohamed, A., M., Altrafi, O.,G., Ismail, M., O., 2014, Relational vs. NoSQL Databases: A Survey, International Journal of Computer and Information Technology, Vol.3, Issue 3, pp. 598-601

Ryan, C., Creech, H., 2012, An Experiment With Social Network Analysis, available at https://www.google.ro/?gws_rd=ssl#q=as+experiment+Network+analysis+software

Vaughan, J., 2014, Big data and cloud computing look for bigger foothold in enterprises, http://searchdatamanagement.techtarget.com//Big-data-and-cloud-computing-look-for-bigger-foothold-in-enterprises

Yahoo Developer Network, 2007, *Hadoop Tutorial*, available at https://developer.yahoo.com/hadoop/tutorial/